DOCUMENT RESUME

ED 093 993 TM 003 849

AUTHOR Brennan, Robert L.
TITLE Some Potential Uses of Decision-Theoretic (Confidence) Testing in the Analysis of Criterion-Referenced Item Data.
PUB DATE [74]
NOTE 43p.

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS *Confidence Testing; *Criterion Referenced Tests; Guessing (Tests); *Item Analysis; Response Style (Tests); Statistical Analysis; Test Reliability; True Scores

ABSTRACT

An attempt is made to explore the use of subjective probabilities in the analysis of item data, especially criterion-referenced item data. Two assumptions are implicit: (1) one wants to obtain a maximum amount of information with respect to an item using a minimum number of subjects; and (2) once the item is validated, it may well be administered in the classical correct/incorrect manner. One way to satisfy these assumptions is to initially administer the unvalidated item to a small number of subjects using confidence testing procedures. Then subjective probabilities can be translated to pseudo-classical item scores which are, at least in theory, guessing-free. Using pseudo-classical scores, a relatively sophisticated item analysis table can be constructed, typical item statistics can be calculated, and a kind of item reliability independent of total test reliability can be assessed. As a useful bridge between subjective probabilities and classical correct/incorrect scores, pseudo-classical scores appear to be of potential use in the analysis of criterion-referenced items. (Author/RC)

ED 093993
TM 003 849

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SOME POTENTIAL USES OF DECISION-THEORETIC (CONFIDENCE) TESTING IN THE ANALYSIS OF CRITERION-REFERENCED ITEM DATA[1]

Robert L. Brennan

S.U.N.Y. at Stony Brook

For norm-referenced testing classical correct/incorrect administration and scoring procedures seem to be reasonably effective and useful. However, norm-referenced tests are usually relatively long; the scores from such tests are often normally distributed; floor and ceiling effects seldom occur in norm-referenced tests; and, most importantly, one is not very much concerned about the precise proportion of items a student can answer correctly--rather, one is concerned about the ability of the test to distinguish among subjects. Each of these characteristics of a norm-referenced test argues directly or indirectly that the classical correct/incorrect procedure is reasonably adequate (or, at least, not grossly inadequate) for many norm-referenced tests.

On the other hand, criterion-referenced tests are usually short; the scores from such tests are often negatively skewed -- even severely so; ceiling effects are very common; and, most importantly, one is fundamentally concerned about accurately estimating the proportion of items to which a student knows the answer (or possibly some other score). This emphasis on accurate estimation of a student's score is especially critical in criterion-referenced testing because there is seldom any external criterion measure for judging validity.

Thus, in criterion-referenced testing it is very important to use every possible means of eliminating random (and systematic) errors of measurement. In particular, it seems to this author that it is important to eliminate (or, at least, be able to estimate the effect of) guessing. Now, it is very clear that, a considerable

---

[1] In this paper "decision-theoretic testing," "confidence testing," and "admissible probability measurement" are all synonymous. The reader is referred to Brennan (1974) for a more complete version of this paper.

amount of student guessing frequently occurs when a student if forced to pick one and only one alternative and the classical correct/incorrect scoring procedure is used; moreover, when the classical procedure is used, it is very difficult, if not impossible, to ascertain the magnitude of the effect of guessing upon student scores.

Furthermore, since criterion-referenced tests are frequently short, it seems desirable to obtain as much informations as possible from each item; yet, using the classical procedure for administering and scoring an item, one merely knows whether or not the student got the item correct. In particular, using the classical procedure one does not obtain information with regard to the relative attractiveness of each alternative for each student. This kind of information can be very useful in determining whether or not to revise a criterion-referenced test item. Thus, the classical procedure somewhat limits the amount of information we obtain with regard to any given criterion-referenced test item.

In short, from a criterion-referenced testing viewpoint, this author feels that the classical procedure for administering and scoring an item has two serious limitations: (a) scores obtained using this procedure incorporate an indeterminable amount of guessing and (b) this procedure provides very little information with regard to any given item especially when relatively small numbers of students take the item. These points imply that when we use the classical procedure for criterion-referenced testing, we may have less than adequate information for determining whether or not a criterion-referenced test item requires revision.

Therefore, it is worthwhile to consider alternatives to the classical procedure. There are a number of points of view from which one could consider different procedures Here we are interested in the ability of the procedure to aid us in item analysis. That is, our goal is to identify a procedure for administering an item that provides us with optimum data for determining whether or not the item needs to be revised; and, if possible, these data should aid us in pinpointing the nature of any difficulties with the item. For this purpose, we consider two potential procedures which we call the "elimination procedure" and the "confidence procedure." We find that the confidence procedure is the better of the two for our purposes.

It should be noted that here we are not concerned about the kinds of scores typically obtained from the elimination and confidence procedures; rather, our primary concern is with the nature and amount of data collected when such procedures are used. Also, we do not assume that once an item is administered using one procedure it will always be administered using that procedure. In fact, when we consider the confidence procedure, the manner in which we interpret the data provides us with a kind of guessing-free estimate of a person's classical score. Thus, once an item has been validated using the confidence procedure, one can administer the item using the classical procedure.

Two Alternatives to the Classical Procedure for Administering Items

Elimination procedure. Coombs et al (1956) suggest a procedure for administering and scoring a test based upon having students eliminate alternatives that they consider to be incorrect. Since a student may eliminate any number of alternatives for any test item, the elimination procedure provides some information about the relative attractiveness of each alternative. However, the information provided is somewhat ambiguous in that, for example, if a student eliminates two alternatives, we so not know whether or not the student feels more uncertain about one alternative than the other.

Also, let us consider the elimination procedure from another point of view. As indicated previously, we are interested in a procedure's ability to provide us with a kind of guessing-free estimate of a person's classical score. Let us call such an estimate a PC1 score, indicating the probability (P) that a person's classical (C) score on an item is unity (1). If we know, for example, that a person guessed randomly on a four-alternative item, then PC1 should be 0.25. The question is, "Can the kind of data collected using the elimination procedure provide us with an adequate basis for estimating a student's PC1 score for an item?"

Suppose, for example, that a student eliminates two alternatives for a four-alternative item. If we could assume that, when forced to pick one and only one alternative, the student would randomly pick one of the two non-eliminated alternatives, then the PC1 score for the student for the item would be 0.50. However, this assumption is not necessarily valid; in fact, one could argue that PC1 might be any value between 0.50 and 1.00.

Thus, it does not appear that the elimination procedure provides an adequate basis for estimating a student's PC1 score for an item. Consequently, if the student were administered the item a large number of times, we don't have a very good basis for estimating the number, or proportion, of times the student would get the item correct under the classical scoring procedure. If the item is administered K times, this proportion should be K·PC1.

Confidence procedure. In confidence testing, one obtains from each student a subjective probability that each alternative of a test item is correct. There are a number of techniques that can be used to obtain these probabilities either directly or indirectly. This author prefers the technique usually called the "star" method in which a student is told to distribute a fixed number of "stars" or points over the alternatives of a test item. For example, students might be told to distribute twelve points over the alternatives of a four-alternative item. The table below indicates some of the ways students might perform this task and the associated (subjective) probabilities.

	No. of Points				Probabilities				
	A*	B	C	D	A*	B	C	D	PC1
S_1	3	3	3	3	.25	.25	.25	.25	.25
S_2	4	4	4	0	.33	.33	.33	.00	.33
S_3	6	6	0	0	.50	.50	.00	.00	.50
S_4	12	0	0	0	1.00	.00	.00	.00	1.00
S_5	5	5	1	1	.42	.42	.08	.08	.50
S_6	5	2	4	1	.42	.17	.33	.08	1.00

The reader interested in a more in-depth discussion of confidence testing can consult de Finetti (1965), Echternacht (1972), Savage (1971), and Shuford et al (1966).[2] A great deal of the literature on confidence testing involves discussion of various procedures for scoring such tests, but this is not our concern in this chapter.

---

[2]Appendix A to Brennan (1974) is a manual for DEC-TEST, a computer program that analyzes confidence test data in great detail. Further, the introduction to this manual provides a description of confidence testing and elimination testing as these procedures are typically used.

Here we are concerned about the nature of the data (i.e., the probabilities) collected for each item and for each student.

Each probability indicates how confident the student is that the particular alternative is the correct answer for the item. Using these probabilities we can obtain PC1 scores from the following rules:

Let M = the magnitude of the highest probability for a particular student for a given item,

A = the number of alternatives for the item,

P(a) = the probability associated with alternative a (a = 1, 2, ..., A), and

* = the correct alternative.

Now,

PC1 = 0 if P(*) $\neq$ M;

PC1 = 1/K if P(*) = M and there are (K-1)other alternatives having P(a) = M; and

PC1 = 1 if P(*) = M and there are not other alternatives having P(a) = M.

See the table on the previous page for examples of PC1 scores. Note, in particular, that the third and fifth students both have PC1 = 0.50 eventhough M = 0.50 for the third student and M = 0.42 for the fifth student.

Thus, PC1 scores are readily available from the subjective probabilities one obtains using the confidence testing procedure. Furthermore, when one uses confidence testing as a procedure to collect data for items, one obtains, for each student, a probability associated with each alternative for each item. Thus, one has a great deal of information for each item -- much more information than if students pick one alternative or eliminate alternatives.

In short, the confidence procedure seems to be superior to the elimination procedure, at least for out purposes here.

Item Analysis Tables from the Confidence Procedure

Conisder the synthetic data for a hypothetical item presented in Table 1. The item has four alternatives, "A" is the correct answer, and the twenty students are partitioned into lower and upper groups of ten students each. The confidence probabilities are indicated for each alternative and for each student. We emphasize that these are synthetic data, and they are not necessarily indicative of a good criterion-referenced test item, we use these data merely to illustrate our discussion.

For each confidence probability in Table 1, there is a pseudo-classical score. A pseudo classical score for an alternative is defined as the probability that a student would pick the alternative if the student were forced to choose one and only one alternative for the item under consideration. Thus, the pseudo-classical score for an item is the pseudo-classical score for the correct alternative; also, the pseudo-classical score for an item is identical to the PC1 score discussed previously.

Using the data in Table 1, one can construct the item analysis tables given by Tables 2 and 3, where Table 2 uses confidence probabilities and Table 3 uses pseudo-classical scores. Both tables present frequency distributions of scores on alternatives, with associated totals, means, and standard deviations. Clearly, Table 2 provides more information, and a somewhat different kind of information than Table 3 ; and, both tables provide much more information than is available from item analysis tables based upon the classical correct/incorrect scoring procedure. This additional information can be quite useful in deciding what (if anything) is wrong with a criterion-refererced test item.

Now, let us summarize a few points implicit in our discussion thus far. We are assuming that once an item is validated it probably will be administered using the classical correct/incorrect scoring procedure. However, in order to validate the item we are suggesting that the evaluator collect confidence probabilities for each alternative, translate these probabilities to pseudo-classical scores for each alternative, and generate the pseudo-classical item analysis table. This table indicates the probability the each student would pick each alternative using the classical correct/incorrect scoring procedure; thus, using this table one can analyze the probable effect of guessing upon the performance of other similar students who take the item using the classical procedure for item administration and scoring. Further,

TABLE 1

Synthetic Data

	Student No.	Confidence Probabilities				Pseudo-classical[a] scores			
		A*	B	C	D	A*	B	C	D
Lower Group	1	.25	.25	.25	.25	.25	.25	.25	.25
	2	.25	.25	.25	.25	.25	.25	.25	.25
	3	.40	.40	.10	.10	.50	.50	.10	.10
	4	1.00	.00	.00	.00	1.00	.00	.00	.00
	5	.30	.20	.30	.20	.50	.00	.50	.00
	6	.50	.50	.00	.00	.50	.50	.00	.00
	7	.30	.30	.10	.30	.33	.33	.00	.33
	8	.20	.70	.00	.10	.00	1.00	.00	.00
	9	.40	.20	.00	.40	.50	.00	.00	.50
	10	.00	1.00	.00	.00	.00	1.00	.00	.00
	Sum-L[b]	3.60	3.80	1.00	1.60	3.83	3.83	1.00	1.33
	Mean-L	.36	.38	.10	.16	.38	.38	.10	.13
	SD-L	.25	.27	.12	.13	.28	.45	.17	.18
Upper Group	11	.25	.25	.25	.25	.25	.25	.25	.25
	12	1.00	.00	.00	.00	1.00	.00	.00	.00
	13	1.00	.00	.00	.00	1.00	.00	.00	.00
	14	.70	.20	.00	.10	1.00	.00	.00	.00
	15	.60	.00	.20	.20	1.00	.00	.00	.00
	16	.50	.50	.00	.00	.50	.50	.00	.00
	17	.40	.50	.00	.10	.00	1.00	.00	.00
	18	.50	.50	.00	.00	.50	.50	.00	.00
	19	.80	.10	.10	.00	1.00	.00	.00	.00
	20	.30	.30	.30	.10	.33	.33	.33	.00
	Sum-U[b]	6.05	2.35	.85	.75	6.58	2.58	.58	.25
	Mean-U	.61	.24	.09	.08	.66	.26	.06	.03
	SD-U	.25	.20	.11	.09	.37	.32	.12	.08
	Sum-T[b]	9.65	6.15	1.85	2.35	10.41	6.41	1.58	1.58
	Mean-T	.48	.31	.09	.12	.52	.32	.08	.08
	SD-T	.28	.25	.12	.12	.12	.34	.15	.15

[a] A pseudo-classical score for an alternative represents the probability that a student would pick the alternative if the student were forced to pick one and only one alternative for the test item.

[b] L, U, and T mean the lower, upper, and total groups, respectively.

TABLE 2

Item Analysis Table Using Confidence Probabilities

Probability Interval	A*			B			C			D		
	Low	Up.	Tot	Low	Up.	Tot	Low	Up.	Tot	Low	Up.	Tot
$0.0 \leq P < 0.1$	1	0	1	1	3	4	5	6	11	3	5	8
$0.1 \leq P < 0.2$	0	0	0	0	1	1	2	1	3	2	3	5
$0.2 \leq P < 0.3$	3	1	4	4	2	6	2	2	4	3	2	5
$0.3 \leq P < 0.4$	2	1	3	1	1	2	1	1	2	1	0	1
$0.4 \leq P < 0.5$	2	1	3	1	0	1	0	0	0	1	0	1
$0.5 \leq P < 0.6$	1	2	3	1	3	4	0	0	0	0	0	0
$0.6 \leq P < 0.7$	0	1	1	0	0	0	0	0	0	0	0	0
$0.7 \leq P < 0.8$	0	1	1	1	0	1	0	0	0	0	0	0
$0.8 \leq P < 0.9$	0	1	1	0	0	0	0	0	0	0	0	0
$0.9 \leq P \leq 1.0$	1	2	3	1	0	1	0	0	0	0	0	0
Total[a]	3.60	6.05	9.65	3.80	2.35	6.15	1.00	.85	1.85	1.60	.75	2.35
Mean[a]	.36	.61	.48	.38	.24	.31	.10	.09	.09	.16	.08	.12
Stan Dev.[a]	.25	.28	.28	.27	.20	.25	.12	.11	.12	.13	.09	.12

[a]These statistics are based upon the actual value of each confidence probability; they are not based upon the midpoint of the probability interval within which the confidence probability lies. See Table 6-1.

TABLE 3

Item Analysis Table Using Pseudo-classical Scores

Pseudo-classical Score	A*			B			C			D		
	Low	Up.	Tot	Low	Up.	Tot	Low	Up.	Tot	Low	Up.	Tot
0.00	2	1	3	3	5	8	7	8	15	6	9	15
0.25	2	1	3	2	1	3	2	1	3	2	1	3
0.33	1	1	2	1	1	2	0	1	1	1	0	1
0.50	4	2	6	2	2	4	1	0	1	1	0	1
1.00	1	5	6	2	1	3	0	0	0	0	0	0
Total[a]	3.83	6.58	10.41	3.83	2.58	6.41	1.00	.58	1.58	1.33	..25	1.58
Mean[a]	.38	.66	.52	.38	.26	.32	.10	.06	.08	.13	.03	.08
Stan Dev[a]	.28	.37	.12	.45	.32	.34	.17	.12	.15	.18	.08	.15

[a]These statistics are based upon the actual value of each pseudo-classical probability.

if one wants a detailed display of the certainty with which students choose any alternative, one can generate the item analysis table based upon the confidence probabilities.

Admittedly, the ideas discussed above require detailed procedures for item administration, scoring, and analysis; however, the additional time and effort required can, I think, be very worthwhile for the process of validating items.

An Application of PC1 Scores in the Classical Test Theory Model

Recall that under the classical test theory model $X = T + E$, where X, T, and E are observed, true, and random error scores, respectively. Now, we have described the PC1 item score for a student as a kind of guessing-free estimate of a person's classical score, and guessing is usually interpreted as one kind of random error. If we assume that guessing is the only, or the principal, kind of random error that concerns us, then a PC1 score is a kind of true score and we can analyze the effect of guessing upon classical scores by using the classical test theory model directly. Thus, in this section we will let

X = 0 or 1 (classical observed score),

T = PC1 item score, and

E = random error due to guessing.

Basic statistics. Note that when one typically uses the classical test theory model, one has observed scores, and one wants to estimate true scores; however, in this case, we already have the true scores, and we must estimate the observed scores. Now, if the item were administered to student i a total of K times we would expect student i to get the item correct $K \cdot T_i$ times, and we would expect student i to get the item incorrect $K \cdot (1-T_i)$ times. Therefore, if N is the total number of subjects

$$\bar{X} = \frac{1}{KN} \sum_{i=1}^{N} K \cdot T_i \qquad (1)$$

$$= \bar{T}$$

$$\text{and } s_X^2 = \frac{1}{KN} \sum_{i=1}^{N} K \cdot T_i - \bar{T}^2$$

$$= \bar{T} - \bar{T}^2$$

$$= \bar{T}(1 - \bar{T}) \quad . \qquad (2)$$

For an example of these statistics see Table 4 which uses the synthetic data presented in Table 1 and assumes, for the sake of illustration, that K = 12.

Table 4 also indicates the error scores associated with each observed score for our synthetic data. The mean and variance of the error scores are given by:

$$\bar{E} = \frac{1}{KN} \sum_{j=1}^{K} \sum_{i=1}^{N} (X_{ij} - T_{ij})$$

$$= \frac{1}{KN} \sum_{i=1}^{N} K \cdot T_i - \frac{1}{N} \sum_{i=1}^{N} T_i$$

$$= \bar{T} - \bar{T}$$

$$= 0 \qquad (3)$$

$$\text{and } s_E^2 = \frac{1}{KN} \sum_{j=1}^{K} \sum_{i=1}^{N} (X_{ij} - T_{ij})^2$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{K} \sum_{j=1}^{K} (X_{ij} - T_{ij})^2 \right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{K} \sum_{j=1}^{K} X_{ij} - \frac{2}{K} \sum_{j=1}^{K} X_{ij} T_{ij} + \frac{1}{K} \sum_{j=1}^{K} T_{ij}^2 \right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left[T_i - \frac{2}{K} (K \cdot T_i^2) + \frac{1}{K} (K \cdot T_i^2) \right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} (T_i - T_i^2)$$

$$= \frac{1}{N} \sum_{i=1}^{N} T_i (1 - T_i) \qquad (4)$$

TABLE 4

Observed, True, and Error Scores

Student	Dist. of Observed Scores		True Scores	Frequency Distribution of Error Scores						
	1	0		-0.50	-0.33	-0.25	0.00	0.50	0.67	0.75
1	3	9	0.25			9				3
2	3	9	0.25			9				3
3	6	6	0.50	6				6		
4	12	0	1.00				12			
5	6	6	0.50	6				6		
6	6	6	0.50	6				6		
7	4	8	0.33		8				4	
8	0	12	0.00				12			
9	6	6	0.50	6				6		
10	0	12	0.00				12			
11	3	9	0.25			9				3
12	12	0	1.00				12			
13	12	0	1.00				12			
14	12	0	1.00				12			
15	12	0	1.00				12			
16	6	6	0.50	6				6		
17	0	12	0.00				12			
18	6	6	0.50	6				6		
19	12	0	1.00				12			
20	4	8	0.33		8				4	

$\bar{X} = 0.521$ $\bar{T} = 0.521$ $\bar{E} = 0.000$

$s_X^2 = 0.249$ $s_T^2 = 0.124$ $s_E^2 = 0.125$

Now, let us demonstrate that $s_X^2 = s_T^2 + s_E^2$.

$$s_T^2 + s_E^2 = \left[\frac{1}{N} \sum_{i=1}^{N} T_i^2 - \bar{T}^2 \right] + \left[\frac{1}{N} \sum_{i=1}^{N} T_i(1 - T_i) \right]$$

$$= \frac{1}{N} \Sigma T_i^2 - \bar{T}^2 + \frac{1}{N} \Sigma T_i - \frac{1}{N} \Sigma T_i^2$$

$$= \bar{T} - \bar{T}^2$$

$$= \bar{T}(1 - \bar{T})$$

$$= s_X^2 \quad .$$

Thus, we have demonstrated that, by interpreting our PC1 scores as true scores we can express the mean and variance of observed scores in terms of the true scores. Furthermore, we have shown that the variance of the observed scores does indeed equal the variance of the true scores plus the variance of the error scores. The mean and variance of the observed, true, and error scores are provided in Table 6-4. For reference now and later, the reader should note that, for our synthetic data

$$\sum_{i=1}^{20} T_i = 10.41 \ ,$$

$$\sum_{i=1}^{20} T_i^2 = 7.9053 \ , \text{ and}$$

$$\sum_{i=1}^{20} T_i^3 = 6.8687 \ .$$

Reliability of a one-item test. Using the above results, we can express the reliability of a one-item test as:

$$r_{11} = s_T^2 / s_X^2$$

$$= \frac{\frac{\Sigma T^2}{N} - \bar{T}^2}{\bar{T}(1 - \bar{T})}$$

$$= \frac{\Sigma T^2 - N \cdot \bar{T}^2}{\Sigma T - N \cdot \bar{T}^2} \qquad (5)$$

For our synthetic data,

$$r_{11} = \frac{0.124}{0.249} = 0.498 \quad .$$

The reader should keep in mind that r_{11} is the proportion of variance in observed scores not due to guessing, whereas $(1 - r_{11})$ is the proportion of variance in observed scores due to guessing. Now, we call r_{11} the reliability of a one-item test; however, if there are random errors operating other than those due to guessing, then r_{11} will be an upper- limit to the "true" reliability of the item.

In order to estimate the reliability of a test consisting of K replications of the item, we can use the Spearman-Brown Prophecy Formula

$$r_{KK} = \frac{Kr_{11}}{1 - (K - 1)r_{11}} \qquad (6)$$

Another way to view the reliability of a one-item test is to ask how many items of a similar nature would have to be administered in order to obtain a given level of reliability. This question can be answered by

re-arranging the terms in the Spearman-Brown Prophecy Formula in order to get

$$K = \frac{r_{KK}(1 - r_{11})}{r_{11}(1 - r_{KK})} , \qquad (7)$$

where, in this case, r_{KK} is the level of reliability desired and K is the number of items necessary to achieve this level of reliability. Using our synthetic data, if we set $r_{KK} = 0.90$, then

$$K = \frac{0.90(1 - 0.498)}{0.498(1 - 0.90)} = 9.072 .$$

One further statistic, of a reliability nature, may be of interest. It can be shown that the probability that a randomly selected student would maintain his or her observed score on L = 2 or 3 administrations of the same item is:

$$P_L = 1 - Ls_E^2 . \qquad (8)$$

For our synthetic data,

$$P_2 = 1 - 2(0.125) = 0.750$$

and $P_3 = 1 - 3(0.125) = 0.635$.

Regression of observed scores on true scores. The standard error of measurement is the square root of the expression in (6.4), which is also equal to

$$s_E = s_X\sqrt{1 - r_{11}} \qquad (9)$$

For our synthetic data,

$$s_E = \sqrt{0.125} = 0.354$$

or $s_E = \sqrt{0.249}\sqrt{1 - 0.498} = 0.354$.

The reader should recall that the standard error of measurement is associated with the regression of observed scores on true scores, as indicated, for our synthetic data, in Figure 6-1. This regression is used to predict observed scores from true scores. As such, this regression can be used to establish a confidence interval around the expected difficulty level of the item, where difficulty level is based on the classical scoring procedure and is merely the proportion of subjects who get an item correct.

Regression of true scores on observed scores. The other regression of interest is the regression of true scores on observed scores. From classical test theory, this regression is:

$$\hat{T} = \bar{T}(1 - r_{11}) + r_{11}X \qquad (10)$$

where $\hat{T}$ is the estimated value of T assuming a linear regression of true on observed scores. The standard deviation of errors about this regression is called the standard error of estimate and denoted s_{est}. For the kind of data considered here, it can be shown that

$$s^2_{est} = \left[\frac{\Sigma T - \Sigma T^2}{N} \right] \left[\frac{N\Sigma T^2 - (\Sigma T)^2}{N\Sigma T - (\Sigma T)^2} \right] \qquad (11)$$

$$= s^2_E \, r_{11}$$

Now, since there are only two possible observed scores for an item (0 and 1) it is also true that

$$s^2_{est} = w_0 s^2_{est(0)} + w_1 s^2_{est(1)} \text{ , where} \qquad (12)$$

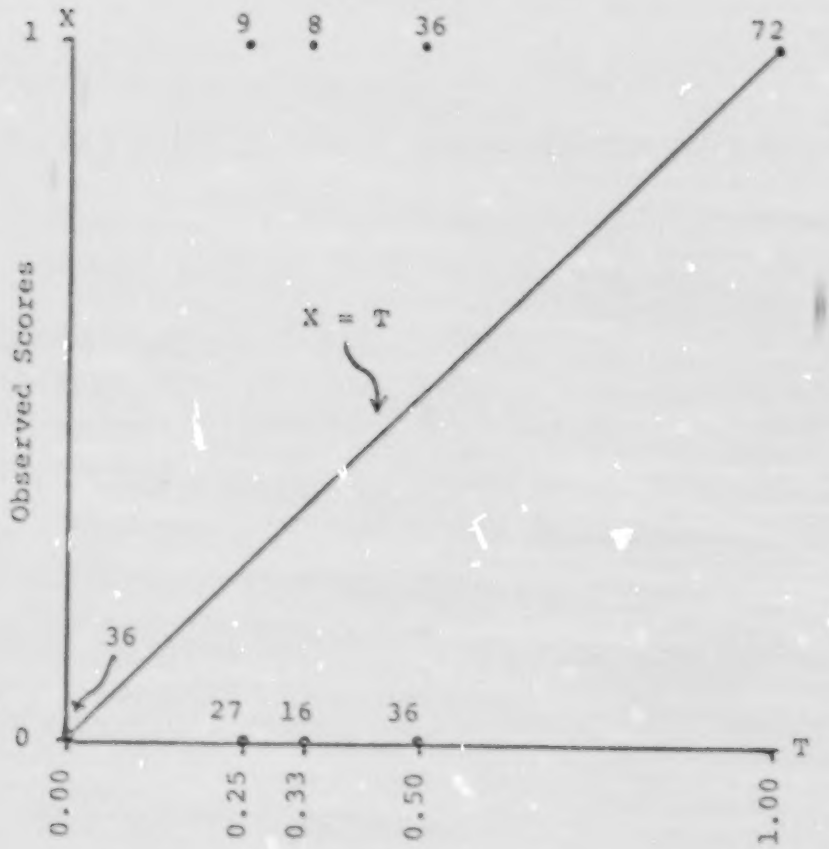
$s^2_{est(0)}$ = the variance of the errors about the regression line when X = 0

$$= \left[\frac{\Sigma T^2 - \Sigma T^3}{N - \Sigma T} \right] - \left[\frac{\Sigma T - \Sigma T^2}{N - \Sigma T} \right]^2 , \qquad (13)$$

$$w_0 = 1 - \bar{T} , \qquad (14)$$

FIGURE 1

Regression of Observed Scores on True Scores



$s_E^2 = 0.125$

$s^2_{est(1)}$ = the variance of the error scores about the regression line when X = 1

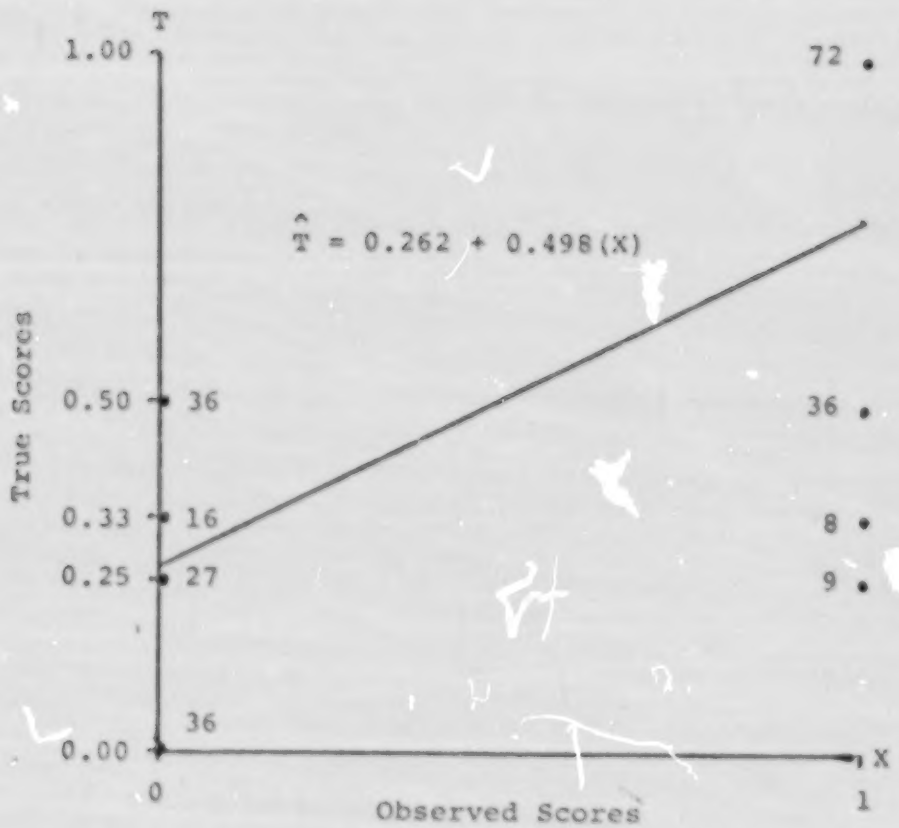
$$= \frac{\Sigma T^3}{\Sigma T} - \left[\frac{\Sigma T^2}{\Sigma T}\right]^2 \quad , \text{ and} \qquad (15)$$

$$w_1 = \bar{T} \quad . \qquad (16)$$

Figure 2 provides, for our synthetic data, the regressions of true scores on observed scores, as well as the values of the statistics indicated in (11), (13), and (15).

FIGURE 2

Regression of True Scores on Observed Scores



$s^2_{est} = 0.062$

$s^2_{est(0)} = 0.040$

$s^2_{est(1)} = 0.083$

Data Analysis[1]

Design for Data Collection. In the fall of 1972 and the spring of 1973 two forms (A and B) of a 25-item criterion-referenced test for a course in educational measurement were administered in both the pre- and posttest mode to 113 students.

In order to understand the design used for administering these tests, the reader will find it useful to refer to the format of Tables 5-7. In these tables the following notation is used:

Factor	Level	Description
A^3	a_1	test administered using SCoRule[2]
A^3	a_2	test administered using "star" technique
B^3	b_1, b_2, b_3, b_4	blocks of subjects
C	c_1	Form A of test
C	c_2	Form B of test
D	d_1	Pretest
D	d_2	Posttest

Also, note that a "." in place of a subscript indicates the mean over all levels of the factor being considered.

---

[1]See Brennan (1974) for a more complete version of the analysis of the data reported here.

[2]The SCoRule is a mechanical device that aids students in assigning subjective probabilities and determining log scores.

[3]Factors A and B should not be confused with forms A and B of the Pretest and the Posttest.

TABLE 5

Means and Standard Deviations -- Pretest and Posttest
VAR(1) = Arithmetic Mean of Item Confidence Scores

	Pretest		Posttest		
	Fm A	Fm B	Fm A	Fm B	N
	c_1d_1	c_2d_1	c_1d_2	c_2d_2	
a_1b_1	.312 .060		.555 .142		21
a_2b_1	.373 .070		.602 .092		10
a_1b_2	.332 .046			.499 .150	19
a_2b_2	.342 .037			.516 .116	9
a_1b_3		.330 .049	.535 .109		17
a_2b_3		.332 .064	.545 .119		9
a_1b_4		.322 .066		.493 .166	20
a_2b_4		.370 .064		.625 .141	8
$a_.b_.$	.333 .057	.333 .061	.556 .120	.518 .153	113

TABLE 6

Means and Standard Deviations -- Pretest and Posttest
VAR(2) = Arithmetic Mean of Item Pseudo-Classical Scores

	Pretest		Posttest		
	Fm A c_1d_1	Fm B c_2d_1	Fm A c_1d_2	Fm B c_2d_2	N
a_1b_1	.360 .086		.658 .132		21
a_2b_1	.424 .081		.700 .069		10
a_1b_2	.378 .071			.569 .156	19
a_2b_2	.403 .046			.573 .108	9
a_1b_3		.367 .07[illegible]	.666 .115		17
a_2b_3		.386 .096	.63[illegible] .143		9
a_1b_4		.357 .078		.626 .150	20
a_2b_4		.427 .091		.737 .097	8
$a_\bullet b_\bullet$	.383 .077	.37[illegible] .082	.664 .118	.614 .148	113

TABLE 7

Means and Standard Deviations -- Pretest and Posttest
VAR (3) = Arithmetic Mean of Classical Scores

	Pretest		Posttest		
	Fm A	Fm B	Fm A	Fm B	N
	c_1d_1	c_2d_1	c_1d_2	c_2d_2	
a_1b_1	.404 .089		.691 .118		21
a_2b_1	.464 .076		.700 .063		10
a_1b_2	.444 .080			.634 .146	19
a_2b_2	.418 .098			.578 .104	9
a_1b_3		.419 .108	.678 .122		17
a_2b_3		.449 .115	.662 .122		9
a_1b_4		.414 .090		.644 .135	20
a_2b_4		.430 .102		.760 .117	8
$a_\bullet b_\bullet$	.429 .087	.424 .100	.685 .110	.646 .139	113

The reader should note several important facts about this design:

(a) If we collapse the levels of the A factor, we see that subjects in the first block received Pretest A and Posttest A, subjects in the second block received Pretest A and Posttest B, subjects in the third block received Pretest B and Posttest A, and subjects in the fourth block received Pretest B and Posttest B. Furthermore, note that subjects were randomly assigned to blocks.

(b) The discussion above indicates that the design is a (balanced) repeated measures design in which half of the available cells are empty; i.e., each subject took one form of the Pretest and one form of the Posttest, and, thus, no subject took both forms of either the Pretest or the Posttest. In the opinion of this author, the constraints incorporated in the design are realistic in that it is often not feasible to obtain repeated measures for equivalent tests in the real world of course development and evaluation.

(c) Although the constraint mentioned above is realistic, it is, nevertheless, somewhat restricting. For example, we cannot obtain direct measures of the equivalence of the two forms of the Pre- and Posttests. Also, when we examine summary statistics for tests and items, these statistics sometimes will be based upon different or partially overlapping samples of subjects.

Another important aspect of the data collection procedure involves the way in which students responded to test items. For each item, each student identified the alternative he or she would pick if forced to pick one and only one alternative; also, each student indirectly reported his or her subjective probabilities for each alternative for each item. Subjects in level a_1 reported actual log scores (range of 0 to 100) for each alternative using a mechanical device called a SCoRule; these log scores were later transformed into subjective probabilities. Students in level a_2 used the twelve-point "star" system for reporting their subjective probabilities.

Summary Statistics for Subjects and Tests. Tables 5-7 report means and standard deviations over tests and persons for:

VAR(1) = Arithmetic mean of item confidence scores; i.e., each subject's score is the arithmetic mean of the subjective probabilities associated with the correct answer to each item. (Range = 0 to 1.)

VAR(2) = Arithmetic mean of item pseudo-classical scores, which are estimated from the subject's subjective probabilities. (Range 0 to 1 .

VAR(3) = Arithmetic mean of classical item scores, which are determined directly from the "pick one" procedure. (Range = 0 to 1.)

Tables 5-7 are presented for the reader who is interested in comparing the different types of scores discussed above. For our purposes, in this chapter, we will concentrate primarily upon pseudo-classical scores. Recall that pseudo-classical scores are estimated classical scores which are determined from the subjective probabilities assigned by subjects to the alternatives of test items. As indicated previously, pseudo-classical scores are much less affected by guessing than are classical scores, one can directly determine a kind of item reliability from pseudo-classical item scores, and pseudo-classical scores are easily interpreted. Pseudo-classical scores, in fact, appear to have most of the advantages and few of the disadvantages of both classical scores and subjective probabilities.

In short, in the opinion of this author, pseudo-classical scores have considerable promise as a basis for validating criterion-referenced, mastery, and possibly norm-referenced test items. It should be noted that once an item has been validated using pseudo-classical scores, one can logically consider subsequently administering and scoring the item using classical procedures.

In the next section we will analyze each of the items that make up both forms of the criterion-referenced Pretest and Posttest. In this section we will continue to emphasize pseudo-classical item scores, although we will, on occasion, report statistics based upon subjective probabilities associated with items and classical item scores.

Item Statistics. Let us review the nature of each of the tests considered here. There are two forms (A and B) of the Pretest and two forms (A and B) of the Posttest. Pretest A and Posttest A are identical, item by item, and the same is true of Pretest B and Posttest B. If we let "i" be a generic item number, then item i on Form A (in both the Pre- and Posttest) is intended to be equivalent to item i on Form B (in both the Pre- and Posttest). In brief, there are two different tests, or sets of items (Form A and Form B) administered at two different times (Pretest and Posttest). Consequently, a complete analysis of item equivalence must consider the issue of equivalence for each item for both the Pretest and Posttest mode.

If we generalize from classical procedures for testing the equivalence of two tests, we would test the equivalence of two items in, say, the Posttest mode, by administering both items to the same set of subjects at the time of the Posttest. Then, if the means and standard deviations of the two items were the same, we could claim that the two items are equivalent, and the correlation between the item scores for the two items could be interpreted as a coefficient of equivalence for the item. However, the design used to collect our data will not permit such a procedure since, as indicated previously, the same subjects never take both forms of an item in either the Pretest or the Posttest mode.

In short, we cannot obtain a direct measure of item equivalence for the two forms of any item given the design for data collection employed here. However, since subjects were randomly assigned to blocks, and since, for the most part, there are no significant differences between block means for the Pre- and Posttests, we can partially consider the statistical issue of item equivalence by examining the differences between Form A and Form B item means and standard deviations. Tables 8 to 10 present the appropriate item statistics when items are scored using subjective (confidence) probabilities, classical scores, and pseudo-classical scores, respectively.

TABLE 8

Item Means and Standard Deviations
Using Confidence Probabilities

	Pretest Means			Posttest Means			Pre Stan Dev's			Post Stan Dev's		
Item	Fm A	Fm B	Diff	Fm A	Fm B	Diff	Fm A	Fm B	FMAX	Fm A	Fm B	FMAX
1	.251	.279	-.028	.289	.311	-.022	.081	.134	2.79***	.255	.215	1.40
2	.418	.242	.176**	.545	.378	.167**	.291	.163	3.21***	.278	.280	1.01
3	.518	.451	.067	.652	.575	.077	.262	.261	1.01	.244	.307	1.59
4	.389	.374	.015	.619	.533	.084	.292	.232	1.58	.320	.272	1.39
5	.310	.303	.007	.400	.407	-.007	.200	.158	1.60	.277	.264	1.11
6	.427	.357	.070	.634	.541	.093	.298	.254	1.38	.292	.286	1.04
7	.229	.297	-.068*	.349	.337	.012	.148	.198	1.78*	.211	.210	1.01
8	.270	.286	-.016	.625	.620	.005	.107	.144	1.82*	.267	.300	1.26
9	.364	.292	.072	.697	.608	.089	.201	.251	1.56	.271	.312	1.32
10	.238	.301	-.063	.251	.426	-.175**	.246	.236	1.08	.341	.328	1.08
11	.433	.586	-.153*	.619	.783	-.164*	.235	.301	1.65	.383	.285	1.81
12	.252	.308	-.056	.717	.635	.082	.124	.181	2.14**	.304	.326	1.15
13	.222	.378	-.156*	.418	.498	-.080	.188	.252	1.78*	.360	.366	1.03
14	.246	.241	.005	.632	.430	.202**	.041	.061	2.24**	.350	.322	1.19
15	.261	.247	.014	.637	.578	.059	.056	.023	5.90***	.330	.321	1.06
16	.270	.310	-.040	.750	.675	.075	.171	.174	1.04	.260	.318	1.49
17	.284	.303	-.019	.669	.623	.046	.129	.170	1.73*	.262	.274	1.09
18	.377	.298	.079*	.674	.474	.200**	.204	.164	1.56	.289	.276	1.10
19	.274	.274	.000	.687	.664	.023	.156	.106	2.15**	.302	.320	1.12
20	.257	.268	-.011	.227	.264	-.037	.179	.199	1.24	.230	.227	1.02
21	.759	.686	.073	.866	.756	.110**	.260	.305	1.38	.170	.265	2.43***
22	.238	.297	-.059	.392	.450	-.058	.153	.214	1.96*	.277	.303	1.20
23	.288	.261	.027	.536	.460	.076	.197	.108	3.37**	.362	.311	1.35
24	.358	.260	.098*	.398	.236	.162**	.195	.181	1.17	.237	.191	1.53
25	.408	.436	-.028	.610	.678	-.068	.219	.211	1.08	.293	.301	1.05
N	59	54		57	56		59	54		57	56	

Note.--All tests are two-tailed.
* $p<.05$ ** $p<.01$ *** $p<.002 = .05/25$

TABLE 9

Item Means and Standard Deviations
Using Classical Scores

	Pretest Means			Posttest Means			Pre Stan Dev's			Post Stan Dev's		
Item	Fm A	Fm B	Diff	Fm A	Fm B	Diff	Fm A	Fm B	FMAX	Fm A	Fm B	FMAX
1	.237	.321	-.084	.228	.357	-.129	.429	.471	1.21	.423	.483	1.30
2	.678	.278	.400***	.772	.554	.218*	.471	.452	1.09	.423	.502	1.40
3	.898	.704	.194**	.947	.732	.215***	.305	.461	2.29**	.225	.447	3.93***
4	.475	.547	-.072	.807	.786	.021	.504	.503	1.00	.398	.414	1.08
5	.559	.556	.003	.561	.536	.025	.501	.502	1.00	.501	.503	1.01
6	.586	.519	.070	.807	.714	.093	.497	.504	1.03	.398	.456	1.31
7	.136	.259	-.123	.456	.393	.063	.345	.442	1.64	.502	.493	1.04
8	.254	.240	.014	.825	.821	.004	.439	.432	1.04	.384	.386	1.01
9	.780	.389	.391***	.895	.786	.109	.418	.492	1.39	.310	.414	1.79
10	.271	.314	-.043	.298	.446	-.148	.448	.469	1.09	.462	.502	1.18
11	.593	.722	-.129	.702	.929	-.227***	.495	.452	1.20	.462	.260	3.15***
12	.153	.566	-.413***	.895	.875	.020	.363	.500	1.90*	.310	.334	1.16
13	.136	.519	-.383***	.404	.589	-.185*	.345	.504	2.13**	.495	.496	1.01
14	.237	.148	.089	.754	.429	.325***	.429	.359	1.43	.434	.499	1.32
15	.237	.185	.052	.667	.750	-.083	.429	.392	1.20	.476	.437	1.18
16	.271	.370	-.099	.912	.839	.073	.448	.487	1.18	.285	.371	1.69
17	.509	.407	.102	.930	.929	.001	.504	.496	1.03	.258	.260	1.02
18	.627	.444	.183	.842	.589	.253**	.488	.502	1.06	.368	.496	1.82*
19	.237	.352	-.115	.807	.786	.021	.429	.482	1.26	.398	.414	1.08
20	.271	.333	-.062	.211	.286	-.075	.448	.476	1.13	.411	.456	1.23
21	.932	.870	.062	1.000	.946	.054	.254	.339	1.79*	.000	.227	----
22	.237	.389	-.152	.474	.518	-.044	.429	.492	1.32	.504	.504	1.00
23	.356	.500	-.144	.597	.607	-.010	.483	.505	1.09	.495	.493	1.01
24	.509	.167	.342***	.456	.107	.349***	.504	.376	1.80*	.502	.312	2.59***
25	.559	.519	.040	.860	.857	.003	.501	.504	1.01	.350	.353	1.02
N	59	54		57	56		59	54		57	56	

Note.--All tests are two-tailed.
* $p<.05$ ** $p<.01$ *** $p<.002 = .05/25$

TABLE 10

Item Means and Standard Deviations
Using Pseudo-Classical Scores

	Pretest Means			Posttest Means			Pre Stan Dev's			Post Stan Dev's		
Item	Fm A	Fm B	Diff	Fm A	Fm B	Diff	Fm A	Fm B	FMAX	Fm A	Fm B	FMAX
1	.253	.323	-.070	.304	.385	-.081	.213	.312	2.14**	.389	.421	1.17
2	.514	.202	.312***	.716	.478	.238**	.438	.294	2.22**	.420	.460	1.20
3	.742	.574	.168*	.889	.707	.182**	.355	.387	1.19	.257	.406	2.51***
4	.448	.421	.027	.809	.729	.050	.414	.397	1.09	.382	.401	1.10
5	.357	.397	-.040	.490	.491	-.001	.321	.358	1.24	.453	.423	1.15
6	.465	.381	.084	.781	.689	.092	.375	.392	1.09	.338	.398	1.39
7	.184	.330	-.146**	.420	.374	.046	.207	.325	2.45***	.409	.375	1.19
8	.280	.299	-.019	.816	.763	.053	.242	.233	1.07	.326	.352	1.17
9	.530	.293	.237***	.851	.731	.120	.377	.343	1.21	.315	.399	1.60
10	.250	.335	-.085	.285	.478	-.193*	.377	.376	1.01	.439	.450	1.05
11	.540	.699	-.159*	.671	.899	-.228***	.358	.389	1.18	.451	.282	2.56***
12	.251	.343	-.092*	.851	.781	.070	.197	.263	1.78*	.313	.357	1.30
13	.213	.443	-.230***	.447	.549	-.102	.270	.375	1.92*	.481	.471	1.05
14	.236	.236	.000	.719	.470	.249**	.139	.090	2.41***	.409	.447	1.19
15	.274	.242	.032	.724	.708	.016	.142	.049	8.29***	.376	.387	1.06
16	.275	.370	-.095	.904	.768	.136*	.333	.358	1.16	.253	.371	2.14**
17	.346	.326	.020	.871	.827	.044	.266	.307	1.33	.290	.310	1.14
18	.504	.360	.144*	.809	.582	.227**	.394	.337	1.37	.358	.419	1.37
19	.257	.279	-.022	.798	.771	.027	.264	.241	1.19	.355	.392	1.22
20	.275	.335	-.060	.228	.281	-.053	.314	.366	1.36	.351	.373	1.12
21	.912	.787	.125*	.991	.915	.076*	.232	.341	2.17**	.066	.235	12.58***
22	.229	.363	-.134*	.436	.479	-.043	.233	.359	2.36***	.389	.409	1.11
23	.309	.269	.040	.588	.539	.049	.276	.167	2.72***	.439	.426	1.06
24	.424	.239	.185**	.444	.147	.297***	.341	.334	1.04	.410	.264	2.41***
25	.517	.546	-.029	.759	.814	-.055	.372	.387	1.08	.378	.322	1.38
N	59	54		57	56		59	54		57	56	

Note.--All tests are two-tailed.
* p<.05 ** p<.01 *** p<.002 = .05/25

TABLE 11

Item Reliabilities

Using Pseudo-Classical Scores

	Pretest			Posttest		
Item	Fm A	Fm B	Diff	Fm A	Fm B	Diff
1	.236	.434	-.198**	.703	.736	-.033
2	.755	.527	.228***	.852	.834	.018
3	.647	.602	.045	.658	.782	-.124*
4	.681	.636	.045	.928	.800	.128*
5	.441	.526	-.085	.807	.704	.103*
6	.556	.641	-.085	.657	.727	-.070
7	.281	.470	-.189**	.675	.590	.085
8	.286	.255	.031	.696	.674	.022
9	.561	.558	.003	.769	.796	-.027
10	.745	.624	.121*	.930	.798	.132*
11	.507	.707	-.200***	.906	.861	.045
12	.203	.302	-.099	.760	.733	.027
13	.428	.560	-.132*	.920	.881	.039
14	.105	.044	.061	.814	.789	.025
15	.100	.013	.087	.696	.712	-.016
16	.547	.541	.006	.725	.759	-.034
17	.307	.422	-.115*	.736	.660	.076
18	.610	.485	.125*	.815	.709	.106*
19	.359	.284	.075	.769	.856	-.087
20	.486	.591	-.105*	.688	.679	.011
21	.659	.682	-.023	.480	.698	-.218***
22	.302	.548	-.246***	.605	.659	-.054
23	.351	.139	.212***	.782	.718	.064
24	.468	.603	-.135*	.669	.546	.123*
25	.545	.594	-.049	.768	.673	.095*
N	59	54		57	56	

* Diff $\geq$ .10
** Diff $\geq$ .15
*** Diff $\geq$ .20

Let us consider Table 10 , which is based upon speudo-classical item scores, in some detail. The means reported can be interpreted in a manner similar to item difficulty levels. The difference between means for the two forms of any item is tested using a t-test for independent samples. The equivalence of item standard deviations is tested using the FMAX statistic, which is the ratio of the larger variance divided by the smaller variance, and which has an F-distribution. Since we are performing multiple tests of significance it is advisable to distribute the a-level (.05) equally over all 25-items; thus, it is advisable to consider a difference or FMAX value to be significant only if $p<.002 = .05/25$.

In addition to comparing means and standard deviations for the two forms ofany item, when we use pseudo-classical scores, we can also compare the item reliabilities discussed previously. These reliabilities are provided in Table 11.

We can summarize the critical information in Tables 10 : and 11 in the following manner.

Item No.	Pretest Differences in:			Posttest Differences in:		
	Mn's	SD's	r's	Mn's	SD's	r's
2	x		x			
3					x	
7		x				
9		x				
11			x	x	x	
13	x					
14		x				
15		x				
21					x	x
22		x	x			
23		x	x			
24				x	x	

In the above table, an "x" appears only if $p<.002$, and the items listed are only those for which at least one pretest or posttest difference is significant at $p<.002$. Clearly there is some evidence that the two forms of some items are not equivalent, for either the pretest mode or the posttest mode or both modes. Note that if two items are equivalent when administered in the pretest mode, this does not guarantee that the items will be equivalent when administered in the posttest mode, and vice-versa.

Table 12 lists the item means (for the two forms of the pre- and posttest) in a format somewhat different from that used in Tables 8-10. Using Table 12 the reader can readily examine the magnitude and direction of the differences among average confidence probabilities, pseudo-classical, and classical scores for each item in each form of each test. It is especially instructive to examine the differences between pseudo-classical and classical scores. Roughly speaking these differences are greater for the pretest than for the posttest; this observation coincides with the fact that pretest item reliabilities are generally lower than posttest item reliabilities.

Table 13 presents correlation coefficients based on the data in Table 12. At least three observations can be made from Table 13:

(a) The correlation between confidence probabilities and classical scores is consistently less than the correlation between pseudo-classical scores and classical scores;

(b) The correlation between confidence probabilities and classical scores is consistently less than the correlation between confidence probabilities and pseudo-classical scores; and

(c) Pretest correlations are consistently less than posttest correlations. This is especially true for the correlations between pseudo-classical and classical scores. This latter observation is to be expected since pretest reliabilities are consistently less than posttest reliabilities.

Summary

This paper should be interpreted as a tentative attempt to explore the use of subjective probabilities (such as those collected when one administers a test in the confidence testing manner) in the analysis of item data, especially criterion-referenced item data.

There are two important assumptions implicit in this paper: (a) one wants to obtain a maximum amount of information with respect to an item using a minimum number of subjects and (b) once the item is validated it may well be administered in the classical correct/incorrect manner. It appears to this author that one way to satisfy these assumptions is to initially administer the unvalidated item to a small number of subjects (say, 20-25) using confidence testing procedures. Then one can translate the subjective probabilities to pseudo-classical item scores which are, at least theoretically,

guessing-free. Using pseudo-classical scores, one can construct a relatively sophisticated item analysis table, calculate typical item statistics, and, in addition, assess a kind of item reliability independent of total test reliability.

Pseudo-classical scores are, in fact, estimates of classical scores; therefore, once the item is validated using pseudo-classical scores, one can subsequently administer and score items in the classical manner. Thus, pseudo-classical scores provide, I think, a useful bridge between subjective probabilities and classical correct/incorrect scores. As such, pseudo-classical scores appear to be of potential use in the analysis of criterion-referenced items.

TABLE 12

Pretest and Posttest Item Means
Using Confidence Probabilities, Pseudo-classical Scores, and Classical Scores[1]

	Pretest Form A			Pretest Form B			Posttest Form A			Posttest Form B		
Item	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1	.251	.253	.237	.279	.323	.321	.289	.304	.228	.311	.385	.357
2	.418	.514	.678	.242	.202	.278	.545	.716	.772	.378	.478	.554
3	.518	.742	.898	.451	.574	.704	.652	.889	.947	.575	.707	.732
4	.389	.448	.475	.374	.421	.547	.619	.809	.807	.533	.729	.786
5	.310	.357	.559	.303	.397	.556	.400	.490	.561	.407	.491	.536
6	.427	.465	.586	.357	.381	.519	.634	.781	.807	.541	.689	.714
7	.229	.184	.136	.297	.330	.259	.349	.420	.456	.337	.374	.393
8	.270	.280	.254	.286	.299	.240	.625	.816	.825	.620	.783	.821
9	.364	.530	.780	.292	.293	.389	.697	.851	.895	.608	.731	.786
10	.238	.250	.271	.301	.335	.314	.251	.285	.298	.426	.478	.446
11	.433	.540	.593	.586	.699	.722	.619	.671	.702	.783	.899	.929
12	.252	.251	.153	.308	.343	.566	.717	.851	.895	.635	.781	.875
13	.222	.313	.136	.378	.443	.519	.418	.447	.404	.498	.549	.589
14	.246	.236	.237	.241	.236	.148	.632	.719	.754	.430	.470	.429
15	.261	.274	.237	.247	.242	.185	.637	.724	.667	.578	.708	.750
16	.270	.275	.271	.310	.370	.370	.750	.904	.912	.675	.768	.839
17	.284	.346	.509	.303	.326	.407	.669	.871	.930	.623	.827	.929
18	.377	.504	.627	.298	.360	.444	.674	.809	.842	.474	.582	.589
19	.274	.257	.237	.274	.279	.352	.687	.798	.807	.664	.771	.786
20	.257	.275	.271	.268	.335	.333	.227	.228	.211	.264	.281	.286
21	.759	.912	.932	.686	.787	.870	.866	.991	1.000	.756	.915	.946
22	.238	.229	.237	.297	.363	.389	.392	.436	.474	.450	.479	.518
23	.288	.309	.356	.261	.269	.500	.536	.588	.597	.460	.539	.607
24	.358	.424	.509	.260	.239	.167	.398	.444	.456	.236	.147	.107
25	.408	.517	.559	.436	.546	.519	.610	.759	.860	.678	.814	.857
N	59	59	59	54	54	54	57	57	57	56	56	56

[1] In this table: (1) = VAR(1) = arithmetic mean of confidence probabilities;
(2) = VAR(2) = arithmetic mean of pseudo-classical scores;
(3) = VAR(3) = arithmetic mean of classical scores.

TABLE 13

CORRELATIONS AMONG THREE DIFFERENT ITEM SCORES

	Pretest Form A			Pretest Form B	
	VAR(1)	VAR(2)		VAR(1)	VAR(2)
VAR(1)			VAR(1)		
VAR(2)	.724		VAR(2)	.853	
VAR(3)	.460	.546	VAR(3)	.498	.668

	Posttest Form A			Posttest Form B	
	VAR(1)	VAR(2)		VAR(1)	VAR(2)
VAR(1)			VAR(1)		
VAR(2)	.849		VAR(2)	.865	
VAR(3)	.772	.908	VAR(3)	.857	.939

NOTE. -- VAR(1) = arithmetic mean of confidence probabilities;
VAR(2) = arithmetic mean of pseudo-classical scores;
VAR(3) = arithmetic mean of classical scores.

BIBLIOGRAPHY

Brennan, R. L. The evaluation of mastery test items. Final report for USOE Grant No. OEG-2-2-2B118. Stony Brook, N.Y. : Department of Education, S.U.N.Y. at Stony Brook, January, 1974.

Coombs, C.H., Milholland, J.E., & Womer, F.B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 13-37.

de Finetti, B. Methods of discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 1965, 13, 87-123.

Echternacht, G.T. The use of confidence testing in objective tests. Review of Educational Research, 1972, 42, 217-236.

Savage, L.J. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association, 1971, 66, 783-801.

Shuford, E.H., Albert, A., & Massengill, H. Admissible probability measurement procedures. Psychometrika, 1966, 31, 125-145.